Laplace's rule of succession in information geometry

Yann Ollivier

When observing data $x_1, \ldots, x_t$ modelled by a probabilistic distribution $p_\theta(x)$, the maximum likelihood (ML) estimator $\theta^{\mathrm{ML}} = \arg\max_\theta \sum_{i=1}^{t} \ln p_\theta(x_i)$ cannot, in general, safely be used to predict x_{t+1}. For instance, for a Bernoulli process, if only "tails" have been observed so far, the probability of "heads" is estimated to 0. (Thus for the standard log-loss scoring rule, this results in infinite loss the first time "heads" appears.)

Bayesian estimators suffer less from this problem, as every value of θ contributes, to some extent, to the Bayesian prediction of x_{t+1} knowing $x_{1:t}$. However, their use can be limited by the need to integrate over parameter space or to use Monte Carlo samples from the posterior distribution.

For Bernoulli distributions, Laplace's famous "add-one" rule of succession (e.g., [CBL06, Grü07]) regularizes θ by adding 1 to the count of "heads" and of "tails" in the observed sequence, thus estimating the Bernoulli parameter p_H by $\hat{p}_H := \frac{n_H+1}{n_H+n_T+2}$ given n_H "heads" and n_T "tails" observations. On the other hand the maximum likelihood estimator is $\frac{n_H}{n_H+n_T}$ so that the two differ at order $O(1/n)$ after $n = n_H + n_T$ observations.

For Bernoulli distributions, Laplace's rule is equivalent to using a uniform Bayesian prior on the Bernoulli parameter [CBL06, Ch. 9.6]. The non-informative Jeffreys prior on the Bernoulli parameter corresponds to Krichevsky and Trofimov's "add-one-half" rule [KT81], namely $\hat{p}_H := \frac{n_H+1/2}{n_H+n_T+1}$. Thus, in this case, some Bayesian predictors have a simple implementation.

We claim (Theorem 1) that for exponential families[1], Bayesian predictors can be approximated by mixing the ML estimator with the *sequential normalized maximum likelihood* (SNML) estimator from universal coding theory [RSKM08, RR08], which is a fully canonical version of Laplace's rule. The weights of this mixture depend on the density of the desired Bayesian prior with respect to the non-informative Jeffreys prior, and are equal to 1/2 for the Jeffreys prior, thus extending Krichevsky and Trofimov's result. The resulting mixture also approximates the "flattened" ML estimator from [KGDR10].

Thus, it is possible to approximate Bayesian predictors without the cost of integrating over θ or sampling from the posterior. The statements below emphasize the special role of the Jeffreys prior and the Fisher information

[1] For simplicity we only state the results with i.i.d. models. However the ideas extend to non-i.i.d. sequences with $p_\theta(x_{t+1}|x_{1:t})$ in an exponential family, e.g., Markov models.

metric. Moreover, the analysis reveals that the direction of the shift from the ML predictor to Bayesian predictors is systematic and given by an intrinsic, information-geometric vector field on statistical manifolds. This could contribute to regularization procedures in statistical learning.

1. Notation and statement. Let $p_\theta(x)\,\mathrm{d}x$ be a family of distributions on a variable x, smoothly parametrized by θ, with density $p_\theta(x)$ with respect to some reference measure $\mathrm{d}x$ (typically $\mathrm{d}x$ is the counting measure for discrete x, or the Lebesgue measure in $\mathbb{R}^d$).

Let $x_1, \ldots, x_t$ be a sequence of observations to be predicted online using p_θ. The maximum likelihood predictor p^{ML} is given by the probability density

$$p^{\mathrm{ML}}(x_{t+1} = y|x_{1:t}) := p_{\theta_t^{\mathrm{ML}}}(y), \qquad \theta_t^{\mathrm{ML}} := \arg\max_\theta \sum_{i=1}^t \ln p_\theta(x_i) \tag{1}$$

assuming this $\arg\max$ is well-defined. Bayesian predictors (e.g., Laplace's rule) usually differ from p^{ML} at order $1/t$.

The *sequential normalized maximum likelihood* predictor ([RSKM08, RR08], see also [TW00]) uses, for each possible value y of x_{t+1}, the parameter $\theta^{\mathrm{ML}+y}$ that would yield the best probability if y had already been observed. Since this increases the probability of every y, it is necessary to renormalize. Define

$$\theta_t^{\mathrm{ML}+y} := \arg\max_\theta \left\{ \ln p_\theta(y) + \sum_{i=1}^t \ln p_\theta(x_i) \right\} \tag{2}$$

as the ML estimator when adding y at position $t+1$. For each y, define the *SNML predictor*[2] for time $t+1$ by the probability density

$$p^{\mathrm{SNML}}(x_{t+1} = y|x_{1:t}) := \frac{1}{Z} p_{\theta_t^{\mathrm{ML}+y}}(y) \tag{3}$$

where Z is a normalizing constant (assuming $Z < \infty$).

For Bernoulli distributions, p^{SNML} coincides with Laplace's "add-one" rule.[3] For other distributions the two may differ[4]: for instance, defining Laplace's rule for continuous-valued x requires choosing a prior distribution on x, whereas the SNML distribution is completely canonical.

[2] This variant of SNML is SNML-1 in [RSKM08] and CNML-3 in [Grü07].

[3] Note that we describe it in a different way. The usual presentation of Laplace's rule is to define $\theta^{\mathrm{Lap}} := \arg\max_\theta \{\ln p_\theta(\text{“heads”}) + \ln p_\theta(\text{“tails”}) + \sum \ln p_\theta(x_i)\}$ and then use θ^{Lap} to predict x_{t+1}. Here we follow the SNML viewpoint and use a different $\theta^{\mathrm{ML}+y}$ for each possible value y of x_{t+1}.

[4] [HB12, BGH$^+$13] contain a characterization of those one-dimensional exponential families for which the variants of NML predictors coincide exactly between themselves and with a Bayesian prior, which is then necessarily the Jeffreys prior. Here Theorem 1 shows that this happens in some approximate sense for *any* exponential family; further relationship between these results is not obvious.

We claim that for exponential families, $\frac{1}{2}(p^{\mathrm{ML}} + p^{\mathrm{SNML}})$ is close to the Bayesian predictor using the Jeffreys prior. This generalizes the “add-one-half” rule.

This extends to any Bayesian prior π by using a *weighted* SNML predictor

$$p^{w\text{-SNML}}(y) := \frac{1}{Z} w(\theta^{\mathrm{ML}+y})\, p_{\theta^{\mathrm{ML}+y}}(y) \tag{4}$$

The weight $w(\theta)$ to be used for a given prior π will depend on the ratio between π and the Jeffreys prior. Recall that the latter is $\pi^{\text{Jeffreys}}(\mathrm{d}\theta) := \sqrt{\det \mathcal{I}(\theta)}\,\mathrm{d}\theta$ where $\mathcal{I}$ is the *Fisher information matrix* of the family (p_θ),

$$\mathcal{I}(\theta) := -\mathbb{E}_{x\sim p_\theta} \partial_\theta^2 \ln p_\theta(x) \tag{5}$$

where ∂_θ^2 stands for the Hessian matrix of a function of θ.

Theorem 1. *Let $p_\theta(x)\,\mathrm{d}x$ be an exponential family of probability distributions, and let π be a Bayesian prior on θ. Then, under suitable regularity assumptions, the Bayesian predictor with prior π knowing $x_{1:t}$ has probability density*

$$\frac{1}{2} p^{\mathrm{ML}}(\cdot|x_{1:t}) + \frac{1}{2} p^{\beta^2\text{-SNML}}(\cdot|x_{1:t}) \tag{6}$$

up to $O(1/t^2)$, where $\beta(\theta)$ is the density of π with respect to the Jeffreys prior, i.e., $\pi(\mathrm{d}\theta) = \beta(\theta)\sqrt{\det \mathcal{I}(\theta)}\,\mathrm{d}\theta$ with $\mathcal{I}$ the Fisher matrix.

More precisely, both under the prior π and under $\frac{1}{2}(p^{\mathrm{ML}} + p^{\beta^2\text{-SNML}})$, the probability density that $x_{t+1} = y$ given $x_{1:t}$ is asymptotically

$$p_{\theta_t^{\mathrm{ML}}}(y)\left(1 + \frac{1}{2t}\left\|\partial_\theta \ln p_\theta(y)\right\|_F^2 + \frac{1}{t}\left\langle \partial_\theta \ln\beta\,, \partial_\theta \ln p_\theta(y)\right\rangle_F - \frac{\dim\Theta}{2t} + O(1/t^2)\right) \tag{7}$$

provided $p_{\theta_t^{\mathrm{ML}}}(y) > 0$, where $\langle \partial_\theta f\,, \partial_\theta g\rangle_F := (\partial_\theta f)^\top \mathcal{I}^{-1}(\theta)\partial_\theta g$ is the Fisher scalar product and $\|\partial_\theta f\|_F^2 = \langle \partial_\theta f\,, \partial_\theta f\rangle_F$ is the Fisher metric norm of $\partial_\theta f$.

For the Jeffreys prior (constant β), this also coincides up to $O(1/t^2)$ with the “flattened” or “squashed” ML predictor from [KGDR10, GK10] with $n_0 = 0$. In particular, the latter is $O(1/t^2)$ close to the Jeffreys prior, and the optimal regret guarantees in [KGDR10] apply to (7). Note that a multiplicative $1 + O(1/t^2)$ difference between predictors results in an $O(1)$ difference on cumulated log-loss regrets.

Regularity assumptions. In most of the article we assume that $p_\theta(x_{t+1}|x_{1:t})$ is a non-degenerate exponential family of probability distributions, with θ belonging to an open set of parameters Θ. The key property we need from exponential families is the existence of a parametrization θ in which $\partial_\theta^2 \ln p_\theta(x) = -\mathcal{I}(\theta)$ for all x and θ: this holds in the natural parametrization for any exponential family (indeed, $p_\theta(x) = \mathrm{e}^{\theta\cdot T(x)}/Z(\theta)$ yields $\partial_\theta \ln p_\theta(x) = T(x) - \partial_\theta \ln Z(\theta)$ so that $\partial_\theta^2 \ln p_\theta(x) = -\partial_\theta^2 \ln Z(\theta)$ for any x).

For simplicity we assume that the space for x is compact, so that to prove $O(1/t^2)$ convergence of distributions over x it is enough to prove $O(1/t^2)$ convergence for each value of x. We assume that the sequence of observations $(x_t)_{t\in\mathbb{N}}$ is an *ineccsi sequence* [Grü07], namely, that for t large enough, the maximum likelihood estimate is well-defined and stays in a compact subset of the parameter space. We also need to assume the same in a Bayesian sense, namely, that for t large enough, the Bayesian maximum a posteriori using prior π is well-defined and stays in a compact subset of Θ. The Bayesian priors are assumed to be smooth with positive densities. On the other hand we do not assume that the Jeffreys prior or the prior π are proper; it is enough that the posterior given the observations is proper, so that the Bayesian predictor at time t is well-defined.

In some parts of the article we do not need p_θ to be an exponential family, but we still assume that the model p_θ is smooth, that there is a well-defined maximum θ_t^{ML} for any $x_{1:t}$ and no other log-likelihood local maxima.

2. Computing the SNML predictor.

We prove Theorem 1 by proving that both predictors are given by (7). Further proofs are gathered at the end of the text.

We first work on p^{SNML}. Here we do not assume that p_θ is an exponential family. Let J_t be the *observed information matrix*, assumed to be positive-definite,

$$J_t(\theta) := -\frac{1}{t}\sum_{i=1}^{t} \partial_\theta^2 \ln p_\theta(x_i) \tag{8}$$

Proposition 2. *Under suitable regularity assumptions, the maximum likelihood update from t to $t+1$ satisfies*

$$\theta_{t+1}^{\mathrm{ML}} = \theta_t^{\mathrm{ML}} + \frac{1}{t} J_t(\theta_t^{\mathrm{ML}})^{-1}\, \partial_\theta \ln p_\theta(x_{t+1}) + O(1/t^2) \tag{9}$$

For exponential families, this update is the natural gradient of $\ln p(x_{t+1})$ with learning rate $1/t$ [Ama98], because $J_t(\theta_t^{\mathrm{ML}}) = \mathcal{I}(\theta_t^{\mathrm{ML}})$, the exact Fisher information matrix. (For exponential families *in the natural parametrization*, $J_t(\theta) = \mathcal{I}(\theta)$ for all θ. But since the Hessian of a function f on a manifold is a well-defined tensor at a critical point of f, it follows that at θ_t^{ML} one has $J_t(\theta_t^{\mathrm{ML}}) = \mathcal{I}(\theta_t^{\mathrm{ML}})$ for *any* parametrization of an exponential family.)

Proposition 3. *Under suitable regularity assumptions,*

$$p^{\mathrm{SNML}}(y|x_{1:t}) = \frac{1}{Z}\, p_{\theta_t^{\mathrm{ML}}}(y)\left(1 + \frac{1}{t}(\partial_\theta \ln p_\theta(y))^\top J_t^{-1}\, \partial_\theta \ln p_\theta(y) + O(1/t^2)\right) \tag{10}$$

provided $p_{\theta_t^{\mathrm{ML}}}(y) > 0$, where J_t is as above and the derivatives are taken at θ_t^{ML}.

Importantly, the normalization constant Z can be computed without having to sum over y explicitly. Indeed (cf. [KGDR10]), by definition of $\mathcal{I}(\theta)$,

$$\mathbb{E}_{y\sim p_\theta}(\partial_\theta \ln p_\theta(y))^\top J_t^{-1}\partial_\theta \ln p_\theta(y) = \operatorname{Tr}(J_t^{-1}\mathcal{I}(\theta)) \tag{11}$$

so that $Z = 1+\frac{1}{t}\operatorname{Tr}(J_t^{-1}\mathcal{I}(\theta_t^{\mathrm{ML}}))+O(1/t^2)$. For exponential families, $J_t = \mathcal{I}$ at θ_t^{ML} so that $Z = 1+\frac{\dim\Theta}{t}+O(1/t^2)$ and

$$p_{\theta_t^{\mathrm{ML}}}(y)\left(1+\frac{1}{t}(\partial_\theta \ln p_\theta(y))^\top \mathcal{I}^{-1}\,\partial_\theta \ln p_\theta(y) - \frac{\dim\Theta}{t}\right) \tag{12}$$

is an $O(1/t^2)$ approximation of $p^{\mathrm{SNML}}(y|x_{1:t})$.

For the weighted SNML distribution $p^{w\text{-SNML}}$, a similar argument yields

$$p^{w\text{-SNML}}(y|x_{1:t}) = \frac{1}{Z}p_{\theta_t^{\mathrm{ML}}}(y)\left(1+\frac{1}{t}(\partial_\theta \ln p_\theta(y))^\top J_t^{-1}\left(\partial_\theta \ln p_\theta(y)+\partial_\theta \ln w(\theta)\right)+O(1/t^2)\right) \tag{13}$$

with $Z = 1+\frac{1}{t}\operatorname{Tr}(J_t^{-1}\mathcal{I}(\theta_t^{\mathrm{ML}}))+O(1/t^2)$ as above. (The $\partial_\theta \ln w$ term does not contribute to Z because $\sum_y p_\theta(y)\partial_\theta \ln p_\theta(y) = 0$.)

Computing $\frac{1}{2}p^{\mathrm{ML}}+\frac{1}{2}p^{w\text{-SNML}}$ with $w(\theta)=\beta(\theta)^2$ in (13), and using that $J_t(\theta^{\mathrm{ML}}) = \mathcal{I}$ for exponential families, proves one half of Theorem 1.

3. Computing the Bayesian posterior. Next, let us establish the asymptotic behavior of the Bayesian posterior. This relies on results from [TK86]. The following proposition may have independent interest.

Proposition 4. *Consider a Bayesian prior $\pi(\mathrm{d}\theta) = \alpha(\theta)\,\mathrm{d}\theta$. Then the posterior mean of a smooth function $f(\theta)$ given data $x_{1:t}$ and prior π is asymptotically*

$$f(\theta_t^{\mathrm{ML}})+\frac{1}{t}(\partial_\theta f)^\top J_t^{-1}\partial_\theta\left(\ln\frac{\alpha}{\sqrt{\det(-\partial_\theta^2 L)}}\right)+\frac{1}{2t}\operatorname{Tr}(J_t^{-1}\partial_\theta^2 f)+O(1/t^2) \tag{14}$$

where $L(\theta) := \frac{1}{t}\ln p_\theta(x_{1:t})$ is the average log-likelihood function, ∂_θ^2 is the Hessian matrix w.r.t. θ, and $J_t := -\partial_\theta^2 L(\theta_t^{\mathrm{ML}})$ is the observed information matrix.

When p_θ is an exponential family in the natural parametrization, for any $x_{1:t}$, $-\partial_\theta^2 L$ is equal to the Fisher matrix $\mathcal{I}$, so that the denominator in the log is the Jeffreys prior $\sqrt{\det\mathcal{I}}$. In particular, for exponential families in natural coordinates, the first term vanishes if the prior π is the Jeffreys prior.

Corollary 5. *Let p_θ be an exponential family. Consider a Bayesian prior $\beta(\theta)\sqrt{\det\mathcal{I}(\theta)}\,\mathrm{d}\theta$ having density β with respect to the Jeffreys prior. Then the posterior probability that $x_{t+1}=y$ knowing $x_{1:t}$ is asymptotically given by (7) as in Theorem 1.*

This proves the second half of Theorem 1.

4. Intrinsic viewpoint. When rewritten in intrinsic Riemannian terms, Proposition 4 emphasizes a systematic discrepancy at order $1/t$ between ML prediction and Bayesian prediction, which is often more "centered" as in Laplace's rule.

This is characterized by a canonical vector field on a statistical manifold indicating the direction of the difference between ML and Bayesian predictors, as follows. In intrinsic terms, the posterior mean (14) in Proposition 4 is[5]

$$f(\theta^{\mathrm{ML}})-\frac{1}{t}(\nabla^2 L)^{-1}\left(\mathrm{d}f, \mathrm{d}\ln\frac{\pi}{\sqrt{\det(-\nabla^2 L)}}\right)-\frac{1}{2t}\operatorname{Tr}\left((\nabla^2 L)^{-1}\nabla^2 f\right)+O(1/t^2) \tag{15}$$

where $L(\theta)=\sum_{i=t}^{t}\ln p_\theta(x_i)$ as above and where ∇^2 is the Riemannian Hessian with respect to any Riemannian metric on θ, for instance the Fisher metric. This follows from a direct Riemannian-geometric computation (e.g., in normal coordinates). In this expression both, the prior $\pi(\mathrm{d}\theta)$ and $\sqrt{\det(-\nabla^2 L)}$ are volume forms on the tangent space so that their ratio is coordinate-independent.[6]

At first order in $1/t$, this is the average of f under a Riemannian Gaussian distribution[7] with covariance matrix $\frac{1}{t}(-\nabla^2 L)^{-1}$, but centered at $\theta^{\mathrm{ML}}-\frac{1}{t}(\nabla^2 L)^{-1}\,\mathrm{d}\ln(\pi/\sqrt{\det(-\nabla^2 L)})$ instead of θ^{ML}.

Thus, if we want to approximate the posterior Bayesian distribution by a Gaussian, there is a systematic shift $\frac{1}{t}V(\theta^{\mathrm{ML}})$ between the ML estimate and the center of the Bayesian posterior, where V is the data-dependent vector field

$$V := -(\nabla^2 L)^{-1}\,\mathrm{d}\ln\left(\pi/\sqrt{\det(-\nabla^2 L)}\right) \tag{16}$$

A particular case is when π is the Jeffreys prior: then

$$V=\frac{1}{2}(\nabla^2 L)^{-1}\,\mathrm{d}\ln\det(-\mathcal{I}^{-1}\nabla^2 L) \tag{17}$$

is an intrinsic vector field defined on any statistical manifold, depending on $x_{1:t}$.

Proposition 6. *When the prior is the Jeffreys prior, the vector V is*

$$V^i=\frac{1}{2}(\nabla_i\nabla_j L)^{-1}(\nabla_k\nabla_l L)^{-1}\,\nabla_j\nabla_k\nabla_l L \tag{18}$$

in Einstein notation, where $L(\theta)=\frac{1}{t}\sum_{s=1}^{t}\ln p_\theta(x_s)$ is the log-likelihood function, and ∇ is the Levi-Civita connection of the Fisher metric.[8]

[5] The equality between (14) and (15) holds only at θ_t^{ML}; the value of (14) is not intrinsic away from θ^{ML}. The equality relies on $\partial_\theta L=0$ at θ^{ML} to cancel curvature contributions.

[6] This is clear when dividing both by the Riemannian volume form $\sqrt{\det g}$: both the prior density $\pi/\sqrt{\det g}$ and $\sqrt{\det(-g^{-1}\nabla^2 L)}$ are intrinsic.

[7] i.e., the image by the exponential map of a Gaussian distribution in a tangent plane.

[8] Note that $\nabla_j\nabla_k\nabla_l L$ is not fully symmetric. Still it is symmetric at θ^{ML}, because the various orderings differ by a curvature term applied to ∇L with vanishes at θ^{ML}.

If p_θ is an exponential family with the Jeffreys prior, the value of V at θ^{ML} does not depend on the observations $x_{1:t}$ and is equal to

$$V^i(\theta^{\mathrm{ML}}) = \frac{1}{4}\mathcal{I}^{ij}\mathcal{I}^{kl}T_{jkl} \tag{19}$$

where T is the skewness tensor [AN00, Eq. (2.28)]

$$T_{jkl}(\theta) := \mathbb{E}_{x\sim p_\theta}\frac{\partial \ln p_\theta(x)}{\partial\theta^j}\frac{\partial \ln p_\theta(x)}{\partial\theta^k}\frac{\partial \ln p_\theta(x)}{\partial\theta^l} \tag{20}$$

$V(\theta^{\mathrm{ML}})$ is thus an intrinsic, data-independent vector field for exponential families, which characterizes the discrepancy between maximum likelihood and the "center" of the Jeffreys posterior distribution. Note that V can be computed from log-likelihood derivatives only. This could be useful for regularization of the ML estimator in statistical learning.

5. Proofs (sketch).

Proof of Proposition 2.

Minimization of a Taylor expansion of log-likelihood around θ_t^{ML}. This is justified formally by applying the implicit function theorem to $F\colon (\varepsilon,\theta)\mapsto \partial_\theta\left(\varepsilon\ln p_\theta(x_{t+1}) + \frac{1}{t}\sum_{i=1}^t \ln p_\theta(x_t)\right)$ at point $(0,\theta^{\mathrm{ML}})$. □

Proof of Proposition 3.

Abbreviate $\theta_y := \theta_t^{\mathrm{ML}+y}$. From Proposition 2 we have

$$\theta_y = \theta_t^{\mathrm{ML}} + \frac{1}{t}J_t^{-1}\partial_\theta \ln p_\theta(y) + O(1/t^2) \tag{21}$$

and expanding $\ln p_\theta(y)$ around θ_t^{ML} yields $p_{\theta_y}(y) = p_{\theta_t^{\mathrm{ML}}}(y)(1 + (\theta_y - \theta_t^{\mathrm{ML}})^\top\partial_\theta \ln p_\theta(y)) + O((\theta_y-\theta^{\mathrm{ML}})^2)$ and plugging in the value of $\theta_y - \theta_t^{\mathrm{ML}}$ yields the result. □

Proof of Proposition 4.

The posterior mean is $(\int f(\theta)\alpha(\theta)p_\theta(x_{1:t})\,\mathrm{d}\theta)/(\int\alpha(\theta)p_\theta(x_{1:t})\,\mathrm{d}\theta)$. From [TK86], if $L_1(\theta) = \frac{1}{t}\ln p_\theta(x_{1:t}) + \frac{1}{t}g_1(\theta)$ and $L_2 = \frac{1}{t}\ln p_\theta(x_{1:t}) + \frac{1}{t}g_2(\theta)$ we have

$$\frac{\int \mathrm{e}^{tL_2(\theta)}\,\mathrm{d}\theta}{\int \mathrm{e}^{tL_1(\theta)}\,\mathrm{d}\theta} = \sqrt{\frac{\det H_1}{\det H_2}}\,\mathrm{e}^{t(L_2(\theta_2)-L_1(\theta_1))}(1+O(1/t^2)) \tag{22}$$

where $\theta_1 = \arg\max L_1$, $\theta_2 = \arg\max L_2$, and H_1 and H_2 are the Hessian matrices of $-L_1$ and $-L_2$ at θ_1 and θ_2, respectively. Here we have $g_1 = \ln\alpha(\theta)$ and $g_2 = g_1 + \ln f(\theta)$ (assuming f is positive; otherwise, add a constant to f).

From a Taylor expansion of L_1 as in Proposition 2 we find $\theta_1 = \theta_t^{\mathrm{ML}} + \frac{1}{t}J_t^{-1}\partial_\theta g_1(\theta_t^{\mathrm{ML}}) + O(1/t^2)$ and likewise for θ_2. So $\theta_1 - \theta_2 = \frac{1}{t}J_t^{-1}\partial_\theta(g_1 -$

$g_2)(\theta_t^{\mathrm{ML}}) + O(1/t^2)$. Since θ_2 maximizes L_2, a Taylor expansion of L_2 around θ_2 gives

$$L_2(\theta_1) = L_2(\theta_2) - \frac{1}{2}(\theta_1 - \theta_2)^\top H_2(\theta_1 - \theta_2) + O(1/t^3) \tag{23}$$

so that, using $L_2 = L_1 + \frac{1}{t}\ln f$ we find

$$\begin{aligned} L_2(\theta_2) - L_1(\theta_1) &= L_2(\theta_1) - L_1(\theta_1) + \frac{1}{2}(\theta_1 - \theta_2)^\top H_2(\theta_1 - \theta_2) + O(1/t^3) \quad (24) \\ &= \frac{1}{t}\ln f(\theta_1) + \frac{1}{2t^2}(\partial_\theta \ln f)^\top J_t^{-1} H_2 J_t^{-1}\, \partial_\theta \ln f + O(1/t^3) \quad (25) \end{aligned}$$

where the second term is evaluated at θ_t^{ML}. We have $H_2 = J_t + O(1/t)$, so $\exp(t(L_2(\theta_2) - L_1(\theta_1))) = f(\theta_1)(1 + \frac{1}{2t}(\partial_\theta \ln f)^\top J_t^{-1}\, \partial_\theta \ln f + O(1/t^2))$. Meanwhile, by a Taylor expansion of $\ln\det(-\partial_\theta^2 L_2(\theta_2))$ around θ_2,

$$\det H_2 = \det(-\partial_\theta^2 L_2(\theta_2)) = \det(-\partial_\theta^2 L_2(\theta_1))\left(1 + (\theta_2 - \theta_1)^\top \partial_\theta \ln\det(-\partial_\theta^2 L_2) + O(\theta_2 - \theta_1)^2\right) \tag{26}$$

and from $L_2 = L_1 + \frac{1}{t}\ln f$ and $\det(A + \varepsilon B) = \det(A)(1 + \varepsilon \operatorname{Tr}(A^{-1}B) + O(\varepsilon^2))$,

$$\begin{aligned} \det(-\partial_\theta^2 L_2(\theta_1)) &= \det(-\partial_\theta^2 L_1(\theta_1))\left(1 + \frac{1}{t}\operatorname{Tr}\left((\partial_\theta^2 L_1)^{-1}\partial_\theta^2(\ln f)\right) + O(1/t^2)\right) \quad (27) \\ &= (\det H_1)\left(1 - \frac{1}{t}\operatorname{Tr}\left(H_1^{-1}\partial_\theta^2(\ln f)\right) + O(1/t^2)\right) \quad (28) \end{aligned}$$

so, collecting,

$$\sqrt{\frac{\det H_1}{\det H_2}} = 1 - \frac{1}{2}(\theta_2 - \theta_1)^\top \partial_\theta \ln\det(-\partial_\theta^2 L_2) + \frac{1}{2t}\operatorname{Tr}\left(H_1^{-1}\partial_\theta^2(\ln f)\right) + O(1/t^2) \tag{29}$$

but $\theta_2 - \theta_1 = J_t^{-1}\partial_\theta \ln f + O(1/t^2)$, and $L_2 = L + O(1/t)$ and $H_1 = J_t + O(1/t)$, so that

$$\sqrt{\frac{\det H_1}{\det H_2}} = 1 - \frac{1}{2t}(\partial_\theta \ln f)^\top J_t^{-1}\partial_\theta \ln\det(-\partial_\theta^2 L) + \frac{1}{2t}\operatorname{Tr}\left(J_t^{-1}\partial_\theta^2(\ln f)\right) + O(1/t^2) \tag{30}$$

Collecting from (22), expanding $f(\theta_1) = f(\theta_t^{\mathrm{ML}})(1 + \frac{1}{t}(\partial_\theta \ln f)^\top J_t^{-1}\partial_\theta \ln\alpha + O(1/t^2))$, and expanding $\partial_\theta \ln f$ in terms of $\partial_\theta f$ proves Proposition 4. □

Proof of Corollary 5.

Let us work in natural coordinates for an exponential family (indeed, since the statement is intrinsic, it is enough to prove it in some coordinate system). In these coordinates, for any x, $\partial^2_\theta \ln p_\theta(x) = -\mathcal{I}(\theta)$ with $\mathcal{I}$ the Fisher matrix, so that $-\partial^2_\theta L = \mathcal{I}(\theta)$. Apply Proposition 4 to $f(\theta) = p_\theta(y)$, expanding $\partial_\theta f = f \partial_\theta \ln f$ and using $\partial^2_\theta \ln f = -\mathcal{I}(\theta)$. □

Proof of Proposition 6.

The Levi-Civita connection on a Riemannian manifold with metric g satisfies $\nabla_l \ln \det A_i^j = (A^{-1})^i_j \nabla_l A_i^j$ thanks to $\partial \ln \det M = \mathrm{Tr}(M^{-1}\partial M)$ and by expanding ∇A. Applying this to $A_i^j = \mathcal{I}^{jk}\nabla^2_{ki} L$ and using $\nabla\mathcal{I} = 0$ proves the first statement. Moreover, for any function f, *at a critical point of* f, $\nabla_l \nabla_j \nabla_k f = \nabla_l \partial_j \partial_k f - \Gamma^i_{jk} \nabla_l \nabla_i f$ and consequently at a critical point of f, with $H_{ij} = \nabla_i \nabla_j f$,

$$\nabla_l \ln \det(g^{ij} H_{jk}) = (H^{-1})^{ij} \nabla_l \partial_i \partial_j f - (H^{-1})^{jk} \Gamma^i_{jk} H_{il} \tag{31}$$

In the natural parametrization of an exponential family, $-\partial^2 L$ is identically equal to the Fisher metric $\mathcal{I}$. Consequently, $\nabla_l \ln \det(-\mathcal{I}^{ij}\nabla^2_{jk} L) = \mathcal{I}^{ij}\nabla_l \mathcal{I}_{ij} - \mathcal{I}^{jk}\Gamma^i_{jk}\mathcal{I}_{il} = -\mathcal{I}^{jk}\Gamma^i_{jk}\mathcal{I}_{il}$ since $\nabla\mathcal{I} = 0$. So from (17), using $d = \nabla = \partial$ for scalars, and $\nabla^2 L = -\mathcal{I}$ at θ^{ML}, we get in this parametrization

$$V^m = -\frac{1}{2}\mathcal{I}^{ml}\partial_l \ln \det(-\mathcal{I}^{-1}\nabla^2 L) = \frac{1}{2}\mathcal{I}^{ml}\mathcal{I}^{jk}\Gamma^i_{jk}\mathcal{I}_{il} = \frac{1}{2}\mathcal{I}^{jk}\Gamma^m_{jk} \tag{32}$$

The Christoffel symbols Γ in this parametrization can be computed from

$$\begin{aligned} \partial_i \mathcal{I}_{jk}(\theta) &= \partial_i \mathbb{E}_{x\sim p_\theta} \partial_j \ln p_\theta(x) \partial_k \ln p_\theta(x) && (33)\\ &= T_{ijk} - \mathcal{I}_{ij}\mathbb{E}_{x\sim p_\theta}\partial_k \ln p_\theta(x) - \mathcal{I}_{ik}\mathbb{E}_{x\sim p_\theta}\partial_j \ln p_\theta(x) = T_{ijk} && (34)\end{aligned}$$

because $\partial_i \partial_j \ln p_\theta(x) = -\mathcal{I}_{ij}(\theta)$ for any x in this parametrization, and because $\mathbb{E}\partial \ln p_\theta(x) = 0$. So $\Gamma^i_{jk} = \frac{1}{2}\mathcal{I}^{il}T_{jkl}$ in this parametrization. This ends the proof. □

Acknowledgments. I would like to thank Peter Grünwald and the referees for valuable comments and suggestions on this text.

References

[Ama98] Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural Comput.*, 10:251–276, February 1998.

[AN00] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 2000. Translated from the 1993 Japanese original by Daishi Harada.

[BGH+13] Peter Bartlett, Peter Grünwald, Peter Harremoës, Fares Hedayati, and Wojciech Kotlowski. Horizon-independent optimal prediction with log-loss in exponential families. In *Conference on Learning Theory (COLT)*, pages 639–661, 2013.

[CBL06] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

[GK10] Peter Grünwald and Wojciech Kotłowski. Prequential plug-in codes that achieve optimal redundancy rates even if the model is wrong. In *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*, pages 1383–1387. IEEE, 2010.

[Grü07] Peter D. Grünwald. *The minimum description length principle*. MIT Press, 2007.

[HB12] Fares Hedayati and Peter Bartlett. The optimality of Jeffreys prior for online density estimation and the asymptotic normality of maximum likelihood estimators. In *Conference on Learning Theory (COLT)*, 2012.

[KGDR10] Wojciech Kotłowski, Peter Grünwald, and Steven De Rooij. Following the flattened leader. In *Conference on Learning Theory (COLT)*, pages 106–118. Citeseer, 2010.

[KT81] R. Krichevsky and V. Trofimov. The performance of universal encoding. *Information Theory, IEEE Transactions on*, 27(2):199–207, 1981.

[RR08] Teemu Roos and Jorma Rissanen. On sequentially normalized maximum likelihood models. In *Proc. 1st Workshop on Information Theoretic Methods in Science and Engineering (WITMSE-2008)*, 2008.

[RSKM08] Teemu Roos, Tomi Silander, Petri Kontkanen, and P. Myllymäki. Bayesian network structure learning using factorized NML universal models. In *Information Theory and Applications Workshop, 2008*, pages 272–276. IEEE, 2008.

[TK86] Luke Tierney and Joseph B Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86, 1986.

[TW00] Eiji Takimoto and Manfred K Warmuth. The last-step minimax algorithm. In *Proceedings of the 11th International Conference on Algorithmic Learning Theory*, pages 279–290. Springer, 2000.